Qualitätsgesteuerte Anfragebearbeitung für Integrierte Informationssysteme

Felix Naumann
Humboldt Universität zu Berlin
Lehrstuhl Datenbanken und Informationssysteme
felix@almaden.ibm.com

Abstract: Die jüngste Entwicklung von zentralisierten Datenbankmanagementsystemen zu Informationssystemen, die verteilte, autonome Datenquellen integrieren, hat zu einer Verschiebung der Forschungsrichtung von der traditionellen *Anfrageoptimierung* zum Bereich der *Anfrageplanung* geführt. Anfrageplanung untersucht das Problem, Pläne zur Anfrageausführung über verteilte, heterogene, überlappende und autonome Datenquellen zu finden. Das wichtigste Unterscheidungsmerkmal der unterschiedlichen Ausführungsstrategien ist nicht mehr—wie für Datenbanksysteme—die Antwortzeit, sondern die *Informationsqualität* des Resultats. Die vorliegende Dissertation untersucht die Verwendung von Qualitätskriterien zur Beantwortung von Nutzeranfragen an integrierte Informationssysteme. Die Arbeit definiert Qualitätskriterien und analysiert, wie sie bewertet werden können. Der wichtigste Teil der Arbeit zeigt, wie diese qualitativen Bewertungen verwendet werden können, um die Qualität der Anfrageresultate zu erhöhen und um die Effizienz von Anfrageplanungsalgorithmen erheblich zu steigern.

1 Web-basierte Informationssysteme

Informationen im Word Wide Web sind verteilt und heterogen, die Informationsquellen sind autonom und verwenden unterschiedliche Schnittstellen, Datenmodelle etc. Um Nutzern die Informationen des Web zu erschließen, werden integrierte Web-basierte Informationssysteme entwickelt. Solche Systeme bieten dem Nutzer eine einheitliche und integrierende Schnittstelle für viele Informationsquellen.
Eine mögliche Architektur zur Informationsintegration ist die Mediator-Wrapper-Architektur (Abb. 1) nach Wiederhold [Wie92]. Ein Mediator hat die Aufgabe, Nutzeranfragen an ein globales Datenschema zu beantworten. Eine Menge autonomer Datenquellen liefert die Daten zu diesem Schema. Für jede dieser Datenquellen steht eine Wrapper-Komponente zur Verfügung, die die Heterogenität der Quelle versteckt, um dem Mediator einen einheitlichen Zugriff zu ermöglichen. Ein Wrapper nimmt Anfragen des Mediators entgegen, übersetzt sie in ein für die Quelle verständliches Format und schickt die übersetzte Anfrage an die Quelle. Darauf nimmt der Wrapper die Resultate entgegen, übersetzt sie wiederum in das Datenmodell des globalen Schemas und reicht das übersetzte Resultat an den Mediator weiter. Dort werden die Ergebnisse integriert und dem Nutzer präsentiert.

In unserer Methode zur qualitätsgesteuerten Anfragebearbeitung bewerten wir jede Quelle gemäß einer Menge von Qualitätskriterien wie Vollständigkeit, Verständlichkeit oder Genauigkeit und stellen Algorithmen vor, die schnell qualitativ gute oder sogar optimale Anfragepläne finden, d.h. Bearbeitungsstrategien, die Resultate höchster Gesamtqualität produzieren. Ein Anwendungsszenario eines solchen Systems ist eine Metasuchmachine, die existierende Suchmaschinen als autonome Quellen integriert. Andere Beispiele sind

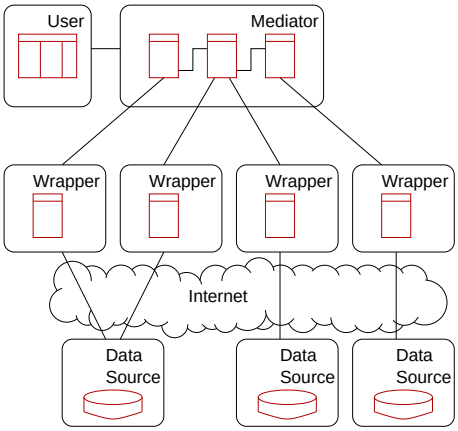

Abbildung 1: Die Mediator-Wrapper Architektur

Aktieninformationsdienste, integrierte Wetterdienste oder verteilte molekularbiologische Datenbanken.

Zur vollständigen Integration von Qualitätsüberlegungen und Anfrageplanung zeigen wir zunächst, wie Datenquellen und Anfragen beschrieben werden können (Abschnitt 2). Insbesondere überlappen sich Quellen in ihren verfügbaren Informationen, so daß Methoden der Integration angewandt werden müssen (Abschnitt 3). Anfragepläne kombinieren einzelne Quellen um dem Nutzer ein Gesamtergebnis liefern zu können (Abschnitt 4). Das Ziel der Arbeit ist es, die besten dieser Pläne zu finden. Zu diesem Zweck definieren wir Informationsqualität als Katalog einzelner Qualitätskriterien (Abschnitt 5) und zeigen Methoden zur Quantifizierung der Kriterien für einzelnen Quellen (Abschnitt 6) und für Pläne über mehrere Quellen (Abschnitt 7). Zur Vervollständigung des Qualitätsmodells beschreiben und untersuchen wir Methoden zur Erstellung von Rangordnungen über mehrere Kriterien (Abschnitt 8). Schließlich beschreiben wir Algorithmen zur effizienten Suche nach den N besten Anfrageplänen (Abschnitt 9).

2 Beschreibung von Datenquellen und Nutzeranfragen

Um einer Nutzeranfrage geeignete Datenquellen zuzuordnen, muß man beide auf kompatible Weise beschreiben. Die Beschreibungen müssen flexibel sein, um Heterogenitäten zu überbrücken und so möglichst viele unterschiedliche Quellen einbinden zu können.

Wir verwenden das Konzept der universellen Relation um sowohl Quellen als auch Anfragen zu beschreiben [MUV84]. Das relationale globale Schema wird in eine universelle Relation transformiert, und Quellen und Nutzeranfragen werden als Sicht auf die universelle Relation beschrieben. Dieser innovative Ansatz erleichtert die Anfrageplanung, ohne die Ausdrucksfähigkeit über Gebühr einzuschränken. Wir stellen das UR-Tableau als

Darstellung der universellen Relation mit den konstituierenden Relationen, Quellen und Anfragen vor (Tab. 1).

	$UR:$	a_1	a_2	$\ldots$	a_{l-1}	a_l
Universal Relation $\longrightarrow$	$R_1:$	✓	✓			
Relations $\longrightarrow$	$\vdots$					
	$R_m:$				✓	✓
	$S_1:$	✓	< 10			
Sources $\longrightarrow$	$\vdots$					
	$S_n:$				✓	='a'
	$Q_1:$	✓			> 20	
User Queries $\longrightarrow$	$\vdots$					
	$Q_r:$		✓			✓

Tabelle 1: Das UR-Tableau

Bei Anfragen an Web-basierte Informationssysteme machen wir Annahmen zur Semantik. Wir nehmen an, daß Nutzer von Informationssystemen im allgemeinen drei Anforderungen (A.1, A.2 und A.3) haben, aber—im Hinblick auf die Problematik Web-basierter, autonomer Systeme—auch zu entsprechenden Zugeständnissen bereit sind (Z.1, Z.2 und Z.3).

A.1 Ein Nutzer erwartet nur korrekte Ergebnisse, d.h. nur Ergebnisse, für die alle Selektionsbedingungen wahr sind. Zum Beispiel erwartet ein Nutzer einer Suchmaschine als Ergebnis *einzig* solche Webseiten, die die Suchwörter enthalten.

Z.1 Ein Nutzer akzeptiert Tupel, deren Attributwerte *nahe* den Selektionsbedingungen sind. Zum Beispiel wird ein Nutzer auf der Suche nach Fahrzeugen unter DM 10.000 auch mit Fahrzeugen einverstanden sein, die für DM 10.500 angeboten werden. Auch dürfen die Resultate einer Suchmaschine Seiten enthalten, in denen Suchwörter im Plural oder als Synonym auftauchen.

A.2 Ein Nutzer erwartet von einem Informationssystem ein *extensional vollständiges* Ergebnis, d.h. es soll alle korrekten Tupel enthalten, die dem System zur Verfügung stehen. Zum Beispiel erwartet ein Nutzer eines Aktieninformationsdienste auf die Anfrage nach Unternehmen mit einem Handelsvolumen über DM 1.000.000 *alle* solche Unternehmen im Ergebnis.

Z.2 Nutzer akzeptieren *extensional unvollständige* Ergebnisse, z.B. bei eingeschränkten Ressourcen. Falls durch solche Einschränkungen nicht das vollständige Ergebnis geliefert werden kann, so sollte das best-mögliche Ergebnis erzielt werden. Bei einer Anfrage an eine Suchmaschine erwartet ein Nutzer nicht immer *alle* Webseiten, die die Suchwörter enthalten—in der Regel genügen etwa die 10 besten Seiten. Ein Ziel dieser Arbeit ist es, die Güte eines solchen Teilergebnisses zu definieren und zu bestimmen.

A.3 Ein Nutzer erwartet ein *intensional vollständiges* Ergebnis, d.h. es sollte alle Attribute der Anfrage enthalten, und kein Attribut sollte `null`-Werte enthalten. Zum Beispiel erwartet ein Nutzer eines Aktieninformationsdienste auf die Anfrage nach Handelsvolumen, Aktienkurs und Umsatz aller DAX-Unternehmen, Tupel, in denen alle diese Daten auch tatsächlich vorhanden sind.

Z.3 Nutzer akzeptieren *intensional unvollständige* Ergebnisse, also Ergebnisse mit fehlenden Attributwerten—eine unvollständige Antwort ist besser als keine Antwort. Ein Nutzer eines Buchinformationsdienstes auf der Suche nach allen Büchern von Stephen King mit Preis und Buchbesprechung wird auch mit einem Ergebnis zufrieden sein, in dem einige Bücher ohne Besprechung vorhanden sind. In solchen Fällen sollte die Tupel zuvorderst gelistet werden, die Werte in allen Attributen haben.

Qualitätsbewertungen und Planungsalgorithmen integrierter Informationssysteme sollten die Anforderungen und Zugeständnisse in sinnvoller Weise unterstützen. D.h. es muß eine Balance zwischen beiden gefunden werden, die höchste Nutzerzufriedenheit gewährleistet. Eine zu strikte Anwendung der Anforderungen kann oft leeren Ergebnisse erzwingen, eine zu lockere Auslegung der Zugeständnisse kann zu übermässig vielen und irrelevanten Ergebnissen führen.

3 Informationsintegration

Autonome Datenquellen überlappen sich extensional, also in den Entitäten, die sie repräsentieren, und intensional, also in den Daten (Attributen) der einzelnen Entitäten. Um eine Nutzeranfrage sinnvoll zu beantworten, muß ein integriertes System deshalb erstens die Überlappungen erkennen und zweitens Datenkonflikte in den sich überlappenden Daten lösen.

Methoden der Objektidentifikation und des *data cleansing* vermögen Daten über gleiche Entitäten zu erkennen [HS98]. Objektidentifikation ist schwierig, da die zugrunde liegenden Daten oft ungenau, unvollständig und inkonsistent sind. Wir verlangen von den Datenquellen zur Identifizierung mehrfach repräsentierter Entitäten den Export eines global eindeutigen und konsistenten Identifikations-Attributs (ID). Ein solches Attribut ist in vielen Web-basierten Anwendungen vorhanden, etwa die URL einer Webseiten oder das Tickersymbol von Aktien.

Nach der Objektidentifikation stehen oft mehrere Repräsentationen der selben Entität zur Verfügung. Zwischen zwei verschiedenen Werten für das gleiche Attribut herrscht ein Datenkonflikt [YM98]. Zur Lösung von Datenkonflikten stellen wir eine allgemeine Konfliktlösungsfunktion vor und geben Beispiele für konkrete Lösungen. Für die eigentliche Integration stellen wir drei neuartige Operatoren vor (siehe Abschnitt 4). Diese verwenden IDs, um Tupel verschiedener Quellen zusammenzuführen, und assoziative Funktionen, um Datenkonflikte aufzulösen.

4 Erstellung von Anfrageplänen

Um Nutzeranfragen gegen die universelle Relation zu beantworten, müssen sie zunächst zu Anfragen gegen die Relationen des globalen Schemas übersetzt werden. Darauf müssen zu der übersetzten Anfrage Anfragepläne gefunden werden. Darin wird festgehalten, welche Quellen verwendet werden, um die Relationen mit Daten zu füllen.
In herkömmlichen Ansätzen werden *alle* Pläne gesucht, die die Anfrage korrekt beantworten [LRO96, Les98]. Die Plaene werden ausgefuehrt und die Ergebnisse der Plaene werden vereinigt. Wir behaupten, daß häufig bereits die besten N Pläne ausreichen, um eine Nutzeranfrage befriedigend zu beantworten, und entwickeln Algorithmen um diese effizient zu finden (Abschnitt 9).
Wir zeigen weiter, daß der herkömmliche Ansatz ungeeignet ist, eine zufriedenstellende Antwort *schnell* zu finden, und stellen eine neues Paradigma der Anfragebearbeitung auf, welches diesen Mangel behebt. Drei neue Integrationsoperatoren erlauben die Suche nach nur noch einem einzigen Plan (Abb. 2). Die Vereinigung der Ergebnisse der einzelnen Pläne im herkömmlichen Paradigma wird nun bereits im Plan selbst, mit Hilfe der outerjoin Operatoren, modelliert.

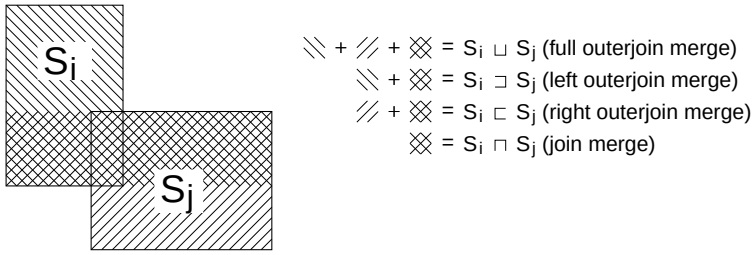

Abbildung 2: Die vier Integrationsoperatoren

5 Definition von Informationsqualität

Wir interpretieren “beste” Pläne und “befriedigende” Antworten als Pläne und Antworten von hoher Informationsqualität. Um Informationsqualität zu bestimmen und zu maximieren, entwickeln wir ein Qualitätsmaß.
Informationsqualität wird durch eine Menge einzelner Qualitätskriterien definiert. Tabelle 2 zeigt eine umfassende Liste solcher Kriterien. Die Liste ist eine Zusammenstellung von Kriterien anderer Projekte, und einiger neuer Kriterien, die speziell für Web-basierte Informationssysteme nötig sind [NR99].
Natürlich sind nicht alle Kriterien in jeder Situation relevant. Vielmehr findet, abhängig von der Anwendung, jeweils eine Auswahl der Kriterien statt. Faktoren, die bei der Auswahl eine Rolle spielen, sind die Anwendungsdomäne, die Art und Präferenzen der vor-

Kategorie	IQ Kriterien	TDQM [WS96]	DWQ [JV97]	Weikum [Wei99]	SCOUG [Bas90]	Chen [CZW98]	Redman [Red96]
Content-related Criteria	Accuracy	✓	✓	✓	✓	✓	✓
	Completeness	✓	✓	✓	✓	✓	✓
	Customer Support				✓		
	Documentation				✓		
	Interpretability	✓	✓				✓
	Relevancy	✓	✓			✓	✓
	Value-Added	✓			✓		
Technical Criteria	Availability	✓	✓	✓	✓		✓
	Latency			✓		✓	
	Price			✓	✓		
	Quality of service			✓	✓		
	Response time			✓		✓	
	Security	✓	✓	✓			
	Timeliness	✓	✓	✓	✓	✓	✓
Intellectual Criteria	Believability	✓	✓	✓	✓		
	Objectivity	✓					
	Reputation	✓	✓				
Instantiation related Criteria	Amount of data	✓				✓	✓
	Repr. conciseness	✓					✓
	Repr. consistency	✓	✓	✓	✓		✓
	Understandability	✓					
	Verifiability			✓			✓

Tabelle 2: Kriterienkatalog für Informationsqualität

aussichtlichen Nutzer, der Anbieter des integrierten Informationssystems, und schließlich die Fähigkeit, die Kriterien zu quantifizieren (siehe Abschnitt 6).

6 Bewertung von Datenquellen nach Qualitätskriterien

Um die Kriterien aus Tab. 2 für Anfrageplanung und Informationsintegration nutzen zu können, müssen sie quantifiziert werden. Wir können drei Quellen solcher Quantifikation identifizieren [NR00]: (i) *Nutzer* können wertvolle Hinweise zur Informationsqualität geben. Viele subjektive Kriterien wie *Verständlichkeit* oder *Ansehen* werden ausschließlich von Nutzer bewertet. In der Regel dienen (Online-)Fragebögen und Feedback-Schleifen als Bewertungsmethode. (ii) Die *Datenquellen* selbst liefern ebenfalls IQ-Bewertungen, wie etwa *Preis* (freiwillig) oder *Vollständigkeit* (unfreiwillig). Solche Bewertungen können zum Teil durch Parsen der Quelle ermittelt werden. (iii) Der *Anfrageprozeß* ergibt IQ-Bewertungen für Kriterien wie *Antwortzeit* oder *Zuverlässigkeit*. Die Werte werden durch vielfaches Befragen der Quelle und Methoden der Statistik ermittelt.

Das Ziel der IQ-Bewertung ist ein IQ-Vektor für jede Informationsquelle. Dieser Vektor enthält für jedes Kriterium eine Bewertung und dient als Grundlage für die Bewertung von Anfrageplänen.

7 Aggregation von Qualitätswerten

Um Nutzeranfragen zu beantworten, ist es oft nötig, Daten von mehr als einer Quelle in einem Plan zu kombinieren. Um die Qualität der kombinierten Antwort (also auch die Qualität des Plans) zu bestimmen, muß man die Qualitätswerte der im Plan beteiligten Quellen aggregieren.
Individuelle *Merge-Funktionen* für Qualitätskriterien fassen zwei Qualitätswerte eines Kriteriums zu einem neuen Wert zusammen [NLF99]. Auf diese Weise aggregieren wir Qualitätswerte entlang des Anfrageplans, um so Qualitätswerte für den Gesamtplan zu erhalten. Abb. 3 zeigt einen solchen Plan mit den drei Informationsquellen S_1, S_2 und S_3, kombiniert über join-Operatoren. Der IQ-Vektor des Plans findet sich unter der Wurzel des Baumes, die Werte sind zusammengesetzt aus denen der einzelnen Quellen.

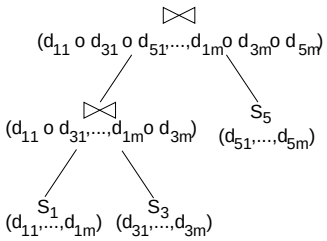

Abbildung 3: Ein Anfrageplan mit IQ-Vektoren und aggregierten IQ-Vektoren

In diesem Zusammenhang schenken wir dem Vollständigkeitskriterium besondere Beachtung. Die Vollständigkeit eines Plans setzt sich zusammen aus dessen Abdeckung (Anzahl der Tupel) und Dichte (Anzahl der nicht-`null` Werte). Wir berechnen Vollständigkeit der Ergebnisse der jeweiligen drei neuen Operatoren und beweisen Eigenschaften, die eine algebraische Optimierung erlauben, ohne die Vollständigkeitswerte zu beeinflussen [NL99, NF00]. Die IQ-Aggregation wird für jeden Anfrageplan vorgenommen. Der Rang der Pläne muß anschließend gemäß ihrer IQ-Vektoren eingeschätzt werden (Abschnitt 8).

8 Erstellung einer Rangordnung gemäß mehrerer Kriterien

Eine Datenquelle oder ein Anfrageplan wird mittels mehrerer Kriterien bewertet. Im allgemeinen haben die Qualitätsmaße der Kriterien unterschiedliche Einheiten und Wertebereiche, außerdem sollen Nutzer die einzelnen Kriterien gewichten können. Um zu entscheiden, welche Quelle die beste oder welcher Plan der beste ist, müssen die einzelnen Werte zu einem Gesamtqualitätswert zusammengefaßt werden.
Wir stellen fünf bekannte Methoden zur Skalierung und Gewichtung vor, und vergleichen sie zur Erstellung von Rangordnungen (Tab. 3). Aufgrund einer Untersuchung allgemeiner Eigenschaften der Methoden und durch Experimente entscheiden wir uns für zwei der Methoden (SAW und DEA) zur Unterstützung der Anfrageplanung [Nau98, NFS98].

	neue beste Quelle	neue schlechteste Quelle	neue ähnliche Quelle	viele Quellen	viele Kriterien
	a / b	a / b	a / b		
SAW	+ / +	+ / +	+ / +	+	+
TOPSIS	+ / +	+ / +	+ / +	+	+
AHP	– / +	– / +	– / –	–	–
ELECTRE	+ / –	+ / –	+ / +	+	+
DEA	+ / –	+ / +	+ / +	+	–

Tabelle 3: Sensibilitätsanalyse der MADM-Methoden (+ bestanden, – nicht bestanden)

9 Effiziente und effektive Anfrageplanung

Algorithmen zur Anfrageplanung sollen effizient zu einer Nutzeranfrage einen oder mehrere Anfragepläne zu finden. Ein Anfrageplan erstreckt sich in der Regel über mehr als eine Quelle.

Wir entwickeln Algorithmen, die mit Hilfe unserer Qualitätsbewertungen Anfrageplanung effizient durchführen und optimale Pläne finden. Wir stellen zwei Ansätze vor, die die besten N Pläne zu einer Anfrage finden. Der erste Ansatz ist eine Erweiterung gängiger Anfrageplanungsalgorithmen um eine qualitative Auswahl von Quellen und Plänen (Abb. 4). Nach einer Vorauswahl der Quellen (Phase 1) wird ein herkömmlichen Planungsalgorithmus verwendet (Phase 2). Die daraus resultierenden Pläne werden qualitativ bewertet (Phase 3), und die besten N zur Ausführung gebracht. Der zweite Ansatz integriert die Qualitätssteuerung mit dem Planungsalgorithmus in einem Branch & Bound Algorithmus—schon während der Planung werden wenig versprechende Teilpläne verworfen. Auf diese Weise verbessern wir nicht nur die Qualität des Resultats, sondern erzielen auch eine dramatische Beschleunigung der Anfrageplanung selbst. [NLF99, LN00]. Auch für das neue Paradigma der Integrationsoperatoren (Abschnitt 4) entwickeln wir Al-

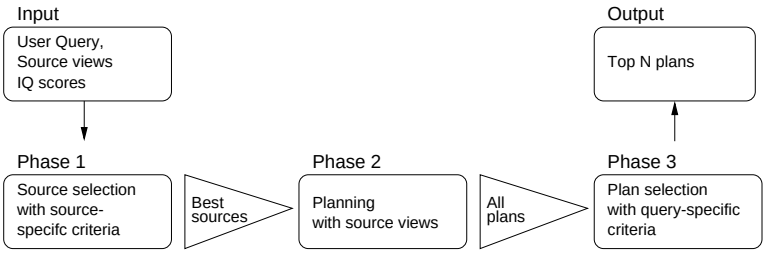

Abbildung 4: Drei-Phasen Anfrageplanung

gorithmen für verschiedene Situationen. Wo wir nicht Optimalität beweisen konnten, testeten wir die Algorithmen erfolgreich in Simulationen [YNGM00, NL00].

10 Zusammenfassung

Der wichtigste Beitrag dieser Dissertation ist die Integration eines Maßes für Informationsqualität mit dem Problem der Anfrageplanung. Die Notwendigkeit und Wichtigkeit von Qualitätsüberlegungen sind in der Forschung hinlänglich bekannt, werden jedoch oft ignoriert; die Integration dieser beiden Bereiche wurde bislang noch nicht vorgenommen. Wir behandeln umfassend alle Aspekte dieser Integration: beginnend mit einem allgemeinen Modell der Anfrageplanung über eine Definition von Informationsqualität und einem Modell, Werte für Qualitätskriterien zu bestimmen, über Vergleichsmethoden um die Qualitätsmaße auszuwerten, bis hin zu Algorithmen, die mit Hilfe dieser Methoden und Maße qualitativ hochwertige Ergebnisse auf Anfragen produzieren.
Daten, die aus autonomen Quellen integriert werden, insbesondere aus WWW Quellen, sind von schlechter Qualität. Diskussionen mit Wissenschaftlern und Praktikern vieler Gebiete haben gezeigt, daß niedrige Informationsqualität eines der drängendsten Probleme moderner Informationssysteme ist. Mit mehr und mehr verfügbaren Daten und mehr und mehr autonomen Quellen wird sich das Problem in Zukunft noch verschärfen. Wir hoffen, daß die Ergebnisse und Methoden dieser Arbeit Eingang in integrierte Informationssysteme finden, um so wieder die Fähigkeit zu erlangen, Nutzern auf effiziente Weise qualitativ hochwertige Daten liefern zu können. Diese Fähigkeit wurde einst verloren: beim Übergang von zentralisierten Datenbanken zu integrierten Informationssystemen.

Literaturverzeichnis

[Bas90] Reva Basch. Measuring the quality of the data: Report on the fourth annual SCOUG retreat. *Database Searcher*, 6(8):18–24, October 1990.

[CZW98] Ying Chen, Qiang Zhu, and Nengbin Wang. Query processing with quality control in the World Wide Web. *World Wide Web*, 1(4):241–255, 1998.

[HS98] Mauricio A. Hernández and Salvatore J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, 2(1):9–37, 1998.

[JV97] M. Jarke and Y. Vassiliou. Data warehouse quality design: A review of the DWQ project. In *Proc. of the Int. Conf. on Information Quality (IQ)*, Cambridge, MA, 1997.

[Les98] Ulf Leser. Combining heterogeneous data sources through query correspondence assertions. In *Workshop on Web Information and Data Management, in conjunction with CIKM'98*, pages 29–32, Washington, D.C., 1998.

[LN00] Ulf Leser and Felix Naumann. Query planning with information quality bounds. In *Proc. of the Int. Conf. on Flexible Query Answering Systems (FQAS)*, Advances in Soft Computing, Warsaw, Poland, 2000. Springer.

[LRO96] Alon Y. Levy, Anand Rajaraman, and Joann J. Ordille. Query-answering algorithms for information agents. In *AAAI National Conf. on Arti£cial Intelligence*, pages 40–47, Portland, ON, 1996.

[MUV84] David Maier, Jeffrey D. Ullman, and Moshe Y. Vardi. On the foundations of the universal relation model. *ACM Transactions on Database Systems (TODS)*, 9(2):283–308, 1984.

[Nau98] Felix Naumann. Data fusion and data quality. In *Proc. of the New Techniques & Technologies for Statistics Seminar (NTTS)*, pages 147–154, Sorrento, Italy, 1998.

[NF00] Felix Naumann and Johann Christoph Freytag. Completeness of information sources. Technical Report 135, Humboldt-Universität zu Berlin, Institut für Informatik, 2000.

[NFS98] Felix Naumann, Johann Christoph Freytag, and Myra Spiliopoulou. Quality-driven source selection using Data Envelopment Analysis. In *Proc. of the Int. Conf. on Information Quality (IQ)*, pages 137–152, Cambridge, MA, 1998.

[NL99] Felix Naumann and Ulf Leser. Density scores for cooperative query answering. In *Proc. of the Workshop Föderierte Datenbanken*, pages 103–116, Berlin, 1999.

[NL00] Felix Naumann and Ulf Leser. Cooperative query answering with density scores. In *Proc. of the Int. Conf. on Management of Data (COMAD)*, Pune, India, 2000.

[NLF99] Felix Naumann, Ulf Leser, and Johann Christoph Freytag. Quality-driven integration of heterogenous information systems. In *Proc. of the Int. Conf. on Very Large Databases (VLDB)*, pages 447–458, Edinburgh, 1999.

[NR99] Felix Naumann and Claudia Rolker. Do metadata models meet IQ requirements? In *Proc. of the Int. Conf. on Information Quality (IQ)*, pages 99–114, Cambridge, MA, 1999.

[NR00] Felix Naumann and Claudia Rolker. Assessment methods for information quality criteria. In *Proc. of the Int. Conf. on Information Quality (IQ)*, Boston, MA, 2000.

[Red96] Thomas C. Redman. *Data Quality for the Information Age*. Artech House, Boston, London, 1996.

[Wei99] Gerhard Weikum. Towards guaranteed quality and dependability of information systems. In *Proc. of the Conf. Datenbanksysteme in Büro, Technik und Wissenschaft (BTW)*, pages 379–409, Freiburg, Germany, 1999.

[Wie92] Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 25(3):38–49, 1992.

[WS96] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal on Management of Information Systems*, 12(4):5–34, 1996.

[YM98] C. Yu and W. Meng. *Principles of database query processing for advanced applications*. Morgan Kaufmann, San Francisco, CA, USA, 1998.

[YNGM00] Ramana Yerneni, Felix Naumann, and Hector Garcia-Molina. Maximizing coverage of mediated web queries. Technical report, Stanford University, CA, 2000. `http://www-db.stanford.edu/~yerneni/pubs/mcmwq.ps`.

Felix Naumann ist geboren am 10.11.1971 in Hamburg, wo er 1990 das Abitur erlangt hat. Von 1990 an studierte er Wirtschaftsmathematik an der Technischen Universität Berlin und schloß 1997 das Studium mit einem Diplom ab. Als Mitglied des Berlin-Brandenburger Graduiertenkollegs "Verteilte Informationssysteme" forschte Felix Naumann von 1997 bis 2000 am Lehrstuhl für Datenbanken und Informationssysteme bei Prof. Johann-Christoph Freytag an der Humboldt Universität zu Berlin, und promovierte in 2000. Seit 2001 ist er als Forscher am IBM Almaden Research Center in San Jose, USA beschäftigt.